International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN (P): 2249-6831; ISSN (E): 2249-7943
Vol. 8, Issue 2, Jun 2018, 69-76

TRANS STELLAR
•Journal Publications •Research Consultancy

THE EFFECTIVENESS OF DATA MINING IN QUERY OPTIMIZATION

HASSANIEN MOHAMMED NAJI[1] & ZAHRAA NAJM ABDULLAH[2]

[1]Department of Telecommunication & Information Technology, Prime Minister's Office, Iraq

[2]Department of Computer Science, College of Science, University of Karbala, Iraq

ABSTRACT

Data mining is a computerized process which is defined as the process of analyzing the big amount of data, and it is a secondary statistical process. Query flocks algorithm is proposed in this research as a data mining technique, which is a generate-and-test model for variety types of patterns, it can be used in facing data mining problems, also it allows the declarative, systematic optimization, and affective processing of a huge set of mining queries. The research specifies the A-priori algorithm, which is the most well-known and primary method in data mining association rules.

The research first defines data mining process and aims from different resources, in addition to clarify the Knowledge discovery Process in Data mining (KDD) and its role in extracting knowledge from data in large database case. Also, the research proposes query flocks framework as a data mining technique here, its algorithm step by step, and its uses. The study results indicated that Data mining technology is using to make the security, scalability, and efficiency better when dealing with a huge amount of data set. It also revealed that there are several limitations of the A-priori algorithm such as; the scanning and checking out time of the data will be too long, and the efficiency will be very low when the database stores a huge amount of data.

KEYWORDS: Data Mining, Query Flocks Algorithm, Efficiency & Optimization

Received: Feb 27, 2018**; Accepted:** Mar 20, 2018**; Published:** May 18, 2018**; Paper Id.:** IJCSEITRJUN20188

Original Article

INTRODUCTION

Managing database systems using query optimization methods seems the most important part of this operation. It is accountable for taking query of the user and checks the whole space of execution policies which tantamount the query of the user, and recurrent the least cost execution plan which can be a thread to the executer. The executer is responsible for carrying out the query (Taylor, 2010). The Cost has the ability to force the plan to change; thus the need of the optimizer is very important to avert poor plans (Taylor, 2010).

In this paper, the usage of data mining in optimizing queries will be discussed. Firstly, data mining will be defined, and its techniques will be mentioned. Secondly, the query optimization will be defined also with discussing its challenges. In addition to clarifying the query flocks to find the final results of using the A-priori algorithm which is produced by it (Yiwu et al., 2008).

DATA MINING

In this section, the definition of data mining will be mentioned. Also, it will clarify the importance in different fields which have the ability to achieve different aims rather than data exploration. Query flocks algorithm is proposed as a data mining technique, which will be discussed in section 3, in addition, to mentioning other data mining techniques.

Data Mining Definition

Data mining is a computerized process which is defined as the process of analyzing the amount of data (usually a large quantity) to create a logical relationship that summarizes data in a new way that is understandable and useful to the data owner. The name "models" is called on the relations and the summary data obtained from the exploration in the data. Data mining usually deals with data obtained for purposes other than data exploration purposes (e. g., a bank's transaction database), which means that the method of data mining never affects the data collection method itself (Clifton and Christopher, 2010). This is one aspect in which data mining is different from statistics so data mining is referred to as a secondary statistical process. The definition also indicates that the amount of data is usually large, and if the amount of data is small, it is preferable to use the normal statistical methods in its analysis (Clifton and Christopher, 2010; Alodibat, 2017).

The widespread and easy availability of information technology has inflated the volume of information in a proactive manner never seen before, making the issue of large data on the Internet controversial, in terms of the usefulness of its existence in this random picture. When talking about massive data, we talk about quantities not it can be imagined from multiple data types and sources with up to hundreds of Terabytes or even Beta-bytes (one bit is followed by 71 zeros). IBM is also talking about a quintile of data every day (Quentillion is the number one followed by 71 zeros). This has led to an increase in the need to develop powerful tools for analyzing data and extracting information and knowledge from them. Traditional methods or statistics cannot deal with this huge quantity, so smart tools are used to process this data (Za'ror, 2015).

Data mining is an important part of the knowledge discovery in database process (KDD) as shown in figure [1], which aims to extract knowledge from data in large database case. So, data mining here helps in identifying what is considered as knowledge, due to the determination of measures, by utilizing a database with the addition to any needed transformations, preprocessing, and subsampling of that database (Fraboni, 2016).

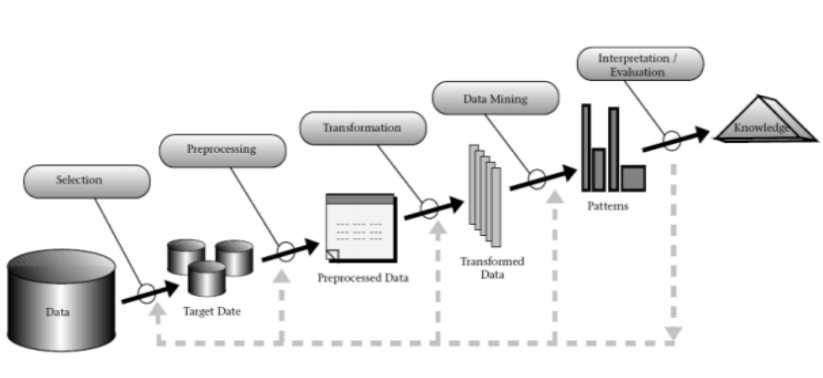

Figure 1: Knowledge Discovery Process in Data Mining (KDD) (Bharati)

Data Mining Techniques

There are several techniques and algorithms in database knowledge discovery such as Clustering, sequence analysis, associations, and Classification rules, Artificial, Regression, Neural Networks, Decision Trees, Intelligence Nearest Neighbor, Genetic Algorithm Association Rules, etc. (Bharati; Imielinski, T. and Mannila, H., 1996).

Data mining process which applied on a structured data has different algorithms which can be applied to finite

kinds of data. Moreover, most methods of data mining are at best loosely coupled with relational database management system (DBMS), so it does not take the advantage of the technology of existing database (Nestorov, 2000). In losing coupling the system of data mining maybe utilizes some of the database functions in addition to data repository system. It reaches the data from the data respiratory, which managed by these performs of data mining on that data. after that, then mining result will be stored in a file or in a specific place in a database or in a data repository (Nestorov, 2000). Thus, a framework, called query flocks is proposed here, in which it declares the formulation of huge types of data mining queries through relational data. Query flock plans method is presented also for systematic efficient optimization processing of every query. A special category of a framework of query flock is that it has the ability to be incorporated in a tightly coupled style with relational database management system (Nestorov, 2000). When using plans of query flock, the query processing abilities can be utilizing relational databases without needing to immolate performance.

QUERY OPTIMIZATION

Database technology has been developed in relational database management systems (RDBMS). In addition to the efficient way to retrieve and store data, it supplies a lot of extra categories like, recoverability, availability, and concurrency controlling. The robust framework of Relational Database (RD) allows the modifications of complex queries, like the online analytical processing (OLAP). Moreover, the effect of finding patterns in relational data can be high, due to the easiness of understanding and exploiting the findings in practice (Nestorov, 2000). One of the techniques in query optimization is a query flock, which is a smart structure for a large data mining classes of problems through relational data.

Query Flocks

Query flocks is a generate-and-test model for variety types of patterns, it can be used in facing data mining problems, query flocks allows the declarative, systematic optimization, and affective processing of a huge set of mining queries. Every mining obstacle is produced as a data-log query with parameters and a candidate status (Yetisgen, 2005).

The most obvious characteristics of query flocks are: (Nestorov, 2000)

- **Declarative Shaping of a Huge Amount of Mining Queries**

In addition to the association-rule mining, there are a lot of variety techniques which deal with specific kinds of data only. For example, different algorithms of association-rule suppose that the data has one relation with just two features. On the other hand, query flocks deal with different kinds of problems of mining.

- **Processing and Optimization Systematically of all Queries**

Structured data mining is specific optimization techniques which applied to some kinds of problems and specific types of data. Systematically mining queries are processed and optimized in query flocks. Plans of a query flock propagate the technique of a-priori for the hugest class of problems of mining

- **Cooperation with Relational DBMS, with Whole Advantage of Founding Abilities**

Most of the present systems of mining are loosely-coupled with RDBMS. So, they leave off the chance to use most of existing abilities of RDBMS. Query flocks have the ability to be cooperated with a relational database.

A-Priori Technique

A-priori (Yiwu et al., 2008) is a set of items used in speeding up searching process. This technique uses a collection of items (S), which is in (c) baskets, any subset of S must be in c baskets at least.

Association rules of data mining is a primary content of data mining exploration at assurance especially is finding the relations of variety database items. A-priori is the most well-known and primary method in data mining association rules (Yiwu et al., 2008). The A-priori precept of the algorithm is to discover the worthy association rules which support and dependability must offset the minimum support and dependability accepted by the user a palm.

A-priori utilize a refined approach which renowned as a level-wise searching; it may frequently browse the dataset to find a matching pattern, then generate a big amount of candidate items in the state of the huge database (Yan-hua and Xia, 2009). For resolving the drawback of A-priori algorithm an improved algorithm will be used.

Agrawal developed the A-priori algorithm in 1994. It is till now the most famous association rule algorithm (Margaret H. and Yongqiao X., no date). A-priori produces the candidate items by gathering the huge items of the prior thread and removing those small subsets are in the prior thread without needing to consider the procedure in the database. By considering huge items of the prior thread, the number of candidate huge items is significantly decreased.

Task pertinent data D is a collection of transactions of a database, every transaction which is expressed by T is a collection of item set (Tid). Suppose item set is I= {I1, I2,.. Im}. Item set which holds k itemset is called a k-itemset. Whether k-item set pleases the Min_sup (minimum support), then k-item set is frequent, and it is indicated by Lk. primarily A-priori algorithm produced a collection of candidates, the candidate k-item set, indicated by Ck. Whether the candidate item pleases the Min_sup (minimum support), so it is a frequent itemset (Yan-hua and Xia, 2009). Figure [2] shows the algorithm of A-priori Technique.

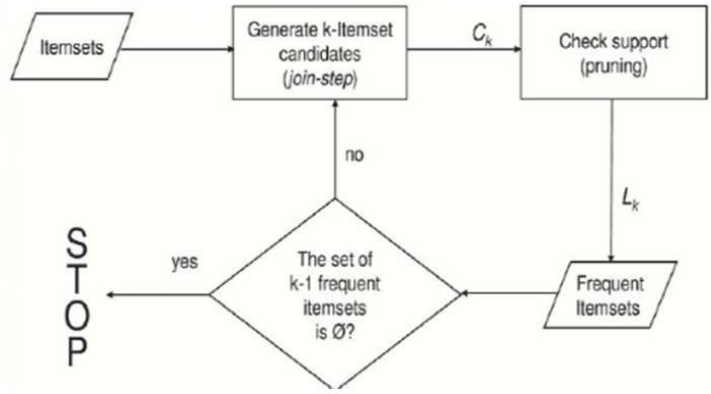

Figure 2: A-Priori Algorithm (Al-Khatib, 2011)

A-priori algorithm is: (Dimitris J., Patrick J., and Amedeo R., 2011)

Let Min_sup is a minimum support threshold and Min_conf is a minimum confidence threshold.

STEP 1: the 1-itemsets candidate, dataset S, C1, and measure how much of appearance of every item is established. The collection of recurrent 1-itemsets, L1 is determined after that, candidate 1-itemsets in C1 having Min_sup.

Check out S again, recurrent 2-itemsets, L2 is determined after that, containing these candidate 2-itemsets in C2 having Min_sup.

STEP 2: frequently check out the dataset S, compare the count of support of every candidate in Ck–1 with Min_sup, and then output Lk–1.

There are two steps used to find the recurrent item sets, which are:

- **Joining Step**: this step is used to find Lk, so, gathering Lk–1 with itself, Ck will be generated.
- **Pruning Step**: in this step, Ck members maybe not frequent. A database must be checked out to determine every Ck candidate count, Lk–1 also must be used to delete a k-item set candidate from Ck, this will be outputted in the determination of Lk.

Figure [3] below shows the A-priori algorithm in an easy way.

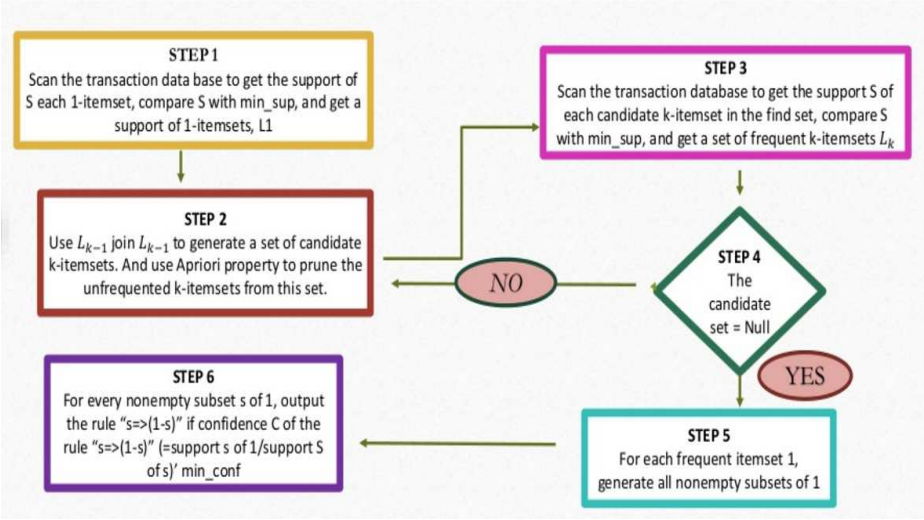

Figure 3: A-Priori Algorithm Steps (International School of Engineering, 2014)

PAPER RESULTS

Data mining technology is using to make the security, scalability, and efficiency better when dealing with a huge amount of data set. This paper shows and clarifies the A-priori algorithm which can be used to increase the efficiency and decrease the consumption time. Query flock produces A-priory algorithm property, which deals with a huge amount of problems. Query flocks analyze a complex data on RDMS.

But, there are some limitations of A-priori algorithms, such as:

- When the database stores a huge amount of data, the scanning and checking out time of the data will be too long, in result the efficiency is very low (Margaret and Yongqiao).
- By increasing the itemsets frequent length, the computing time will be increased (Margaret and Yongqiao).
- A-priori algorithm will generate all frequent item sets candidates, so when frequent itemsets is founded, scanning database is needed frequently. It also will consume more time and resource to achieve one scanning. So it will be not efficient (Yan-hua and Xia, 2009).

FUTURE WORK

The researcher suggests in future studies the selection of other algorithms that contribute in reducing of energy consumed, and contributes in mining the big data, with less complexity than the algorithm used in this study, and without needing to a redundant scanning for the database.

CONCLUSIONS

First, the study defines data mining process and its objectives which needs to the smart algorithm to assist in achieving these aims, in addition, to clarify the role of data mining in a Knowledge discovery Process (KDD) in extracting knowledge from data in large database case. Query flocks technique is proposed here as a data mining technique, which helps in reducing the energy consumption.

After studying the query flock method step by step, it is obvious that it is makes the security, scalability, and efficiency of mining the big data better, in addition to reducing the consumption time. Query flocks have a huge ability in dealing with a huge amount of problems. But it has some of the limitations: scanning and checking out time is directly proportional to the size of data in a database, computing time is directly proportional to frequent length of items, and scanning database is needed frequently when frequent itemsets is founded, so that will increase the consumption time, and decrease the efficiency.

REFERENCES

1. *Al-khatib N. (2011). A-priori algorithm.*
2. *Alodibat, S. (2017). " An overview of the visualization features in open source data mining tools, MECS Journal, 1(1).*
3. *Bharati M. (no date). Data Mining Techniques And Applications.*
4. *Clifton and Christopher (2010). Encyclopedia Britannica: Definition of Data Mining.*
5. *Dimitris J., Patrick J., and Amedeo R. (2011). A Priori Optimization.*
6. *Fraboni E. (2016). Data mining and knowledge discovery.*
7. *Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction*
8. *Imielinski, T. and Mannila, H. (1996). A database perspective on knowledge discovery Communications of the ACM, 39(11):58-64.*
9. *International School of Engineering (2014). A-priori algorithm.*
10. *Jeffrey D. (1998). Query flocks: A generalization of association-rule mining.*
11. *Margaret H. and Yongqiao X. (no date). Survey of Association rule, Department of Computer Science and Engineering, Southern Methodist University, Dallas, Texas.*
12. *Nestorov S. (2000). Data Mining Techniques For Structured And Semi structured Data.*
13. *Taylor R. (2010). Query Optimization for Distributed Database Systems, Master of Computer Science Computing Laboratory University of Oxford.*

14. *Yan-hua W. and Xia F. (2009). The Optimization of A-priori Algorithm Based on Directed Network, Third International Symposium on Intelligent Information Technology Application, IEEE.*
15. *Yetisgen M., Ismail H. (2005). Data mining in deductive databases using query flocks.*
16. *Yiwu X., Yutong L., Chunli W., Mingyu L. (2008). The Optimization and Improvement of the A-priori Algorithm. International Symposium on Intelligent Information Technology Application Workshops, IEEE.*
17. *Za'ror E. (2015). Data mining and economic expectation.*